LLM-Powered Chatbot Assistants

Elevating Shopee's Customer Journey with Generative AI and Intelligent Assistants.

Huiyu Chen | Senior Machine Learning Engineer

SHOPEE

Who We Are

Shopee

The Global E-Commerce Giant

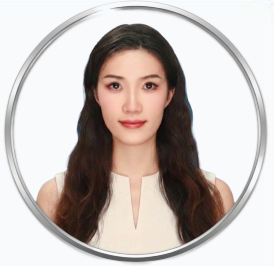
Shopee is the leading e-commerce platform across Southeast Asia and Latin America, connecting millions of buyers and sellers daily.

The Chatbot Team

Our Team's Mission: The Chatbot team is dedicated to building a digital bridge—transforming the Shopee experience from reactive FAQ answering into a proactive, highly intelligent AI-driven interaction at massive scale.

Huiyu Chen

Senior Machine Learning Engineer



Experience: 3+ years specializing in Large Language Models (LLMs) and Recommendation Systems. M.S. in Computer Science from CASIA.

Core Focus: Designing and deploying end-to-end LLM solutions, covering the full lifecycle from RAG architecture and model fine-tuning to high-concurrency deployment.

Today's Agenda

1. The Landscape: E-Commerce Dilemma
2. Core Architecture: 4 AI Pillars
3. Training the Brain: SFT & Alignment
4. Conclusion & Key Takeaways

Today's Agenda

1. The Landscape: E-Commerce Dilemma
2. Core Architecture: 4 AI Pillars
3. Training the Brain: SFT & Alignment
4. Conclusion & Key Takeaways

The Seller's Dilemma

Efficiency vs. Cost

The internet never sleeps. Stores are open 24/7, and questions flood in at 3 AM from halfway across the globe.

The obvious answer—hiring human agents around the clock—gets wildly expensive. It's often a total non-starter for growing businesses.

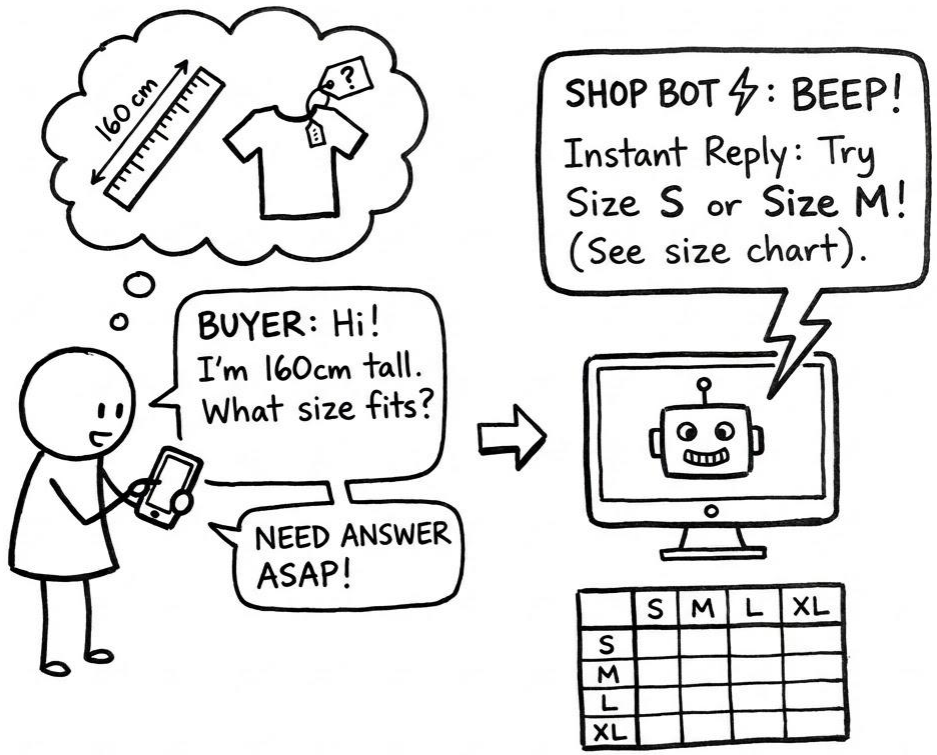
Sellers are desperate to boost efficiency, cut costs, offer 24/7 support, and still build genuine customer connections. It is a massive tall order.

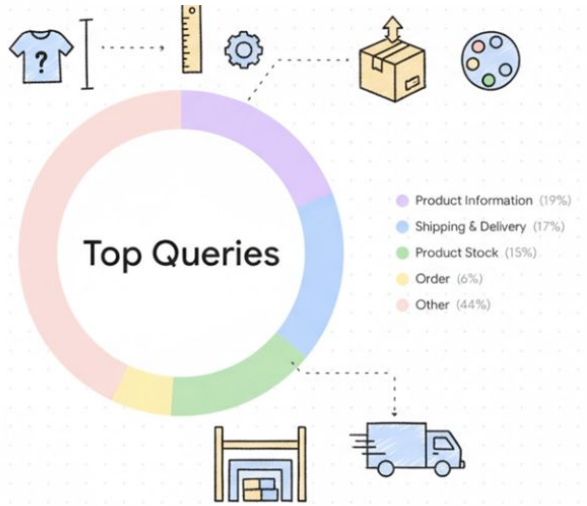
The Buyer's Quest

Confidence & Information

For buyers, it's not about operational costs; it's about confidence. It's a quest for the right information, *right now*.

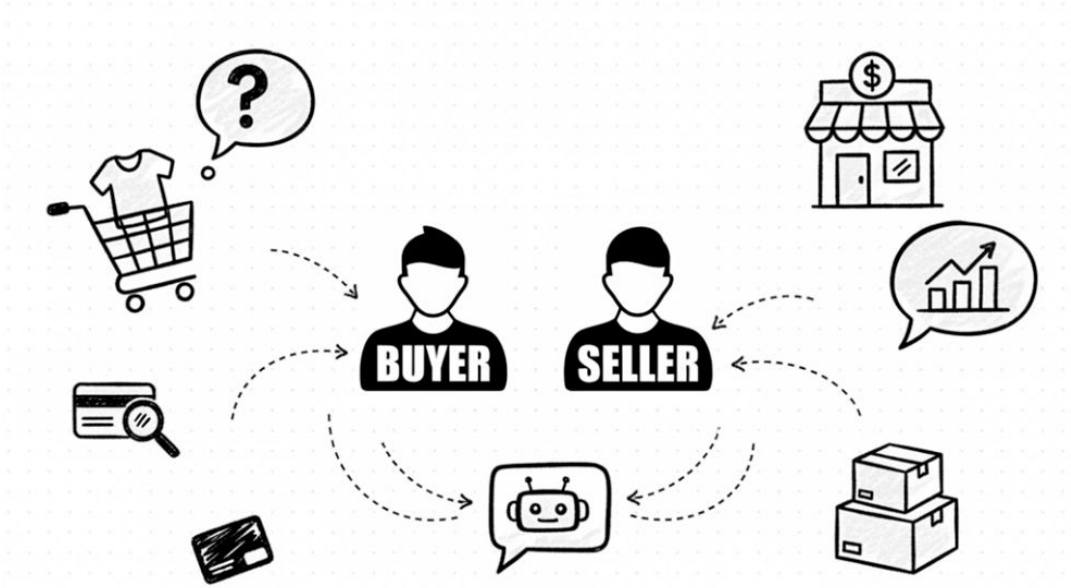
- **The Baseline:** Almost 20% of questions are basic product info (color, size, fit).
- **The Friction:** Closely followed by critical logistical queries—shipping costs and restock dates.
- **The Goal:** Answering these fast turns "I don't know" into "I need this", building the trust required to hit "Add to Cart."

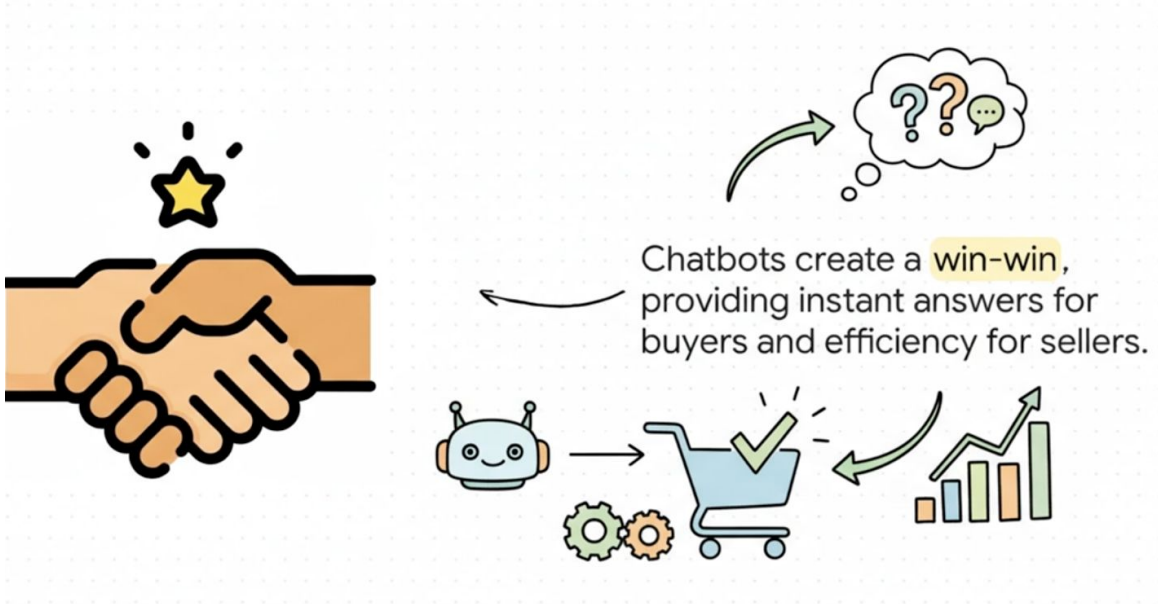




The top questions from buyers focus on basic product and delivery details.

The Digital Bridge





A Mutual Win-Win

You have sellers totally overwhelmed, and buyers full of questions. This creates a huge gap.

The chatbot acts as the perfect digital handshake. For the buyer, it means instant answers and zero frustration. For the seller, it means massive efficiency and a much better shot at closing the sale.

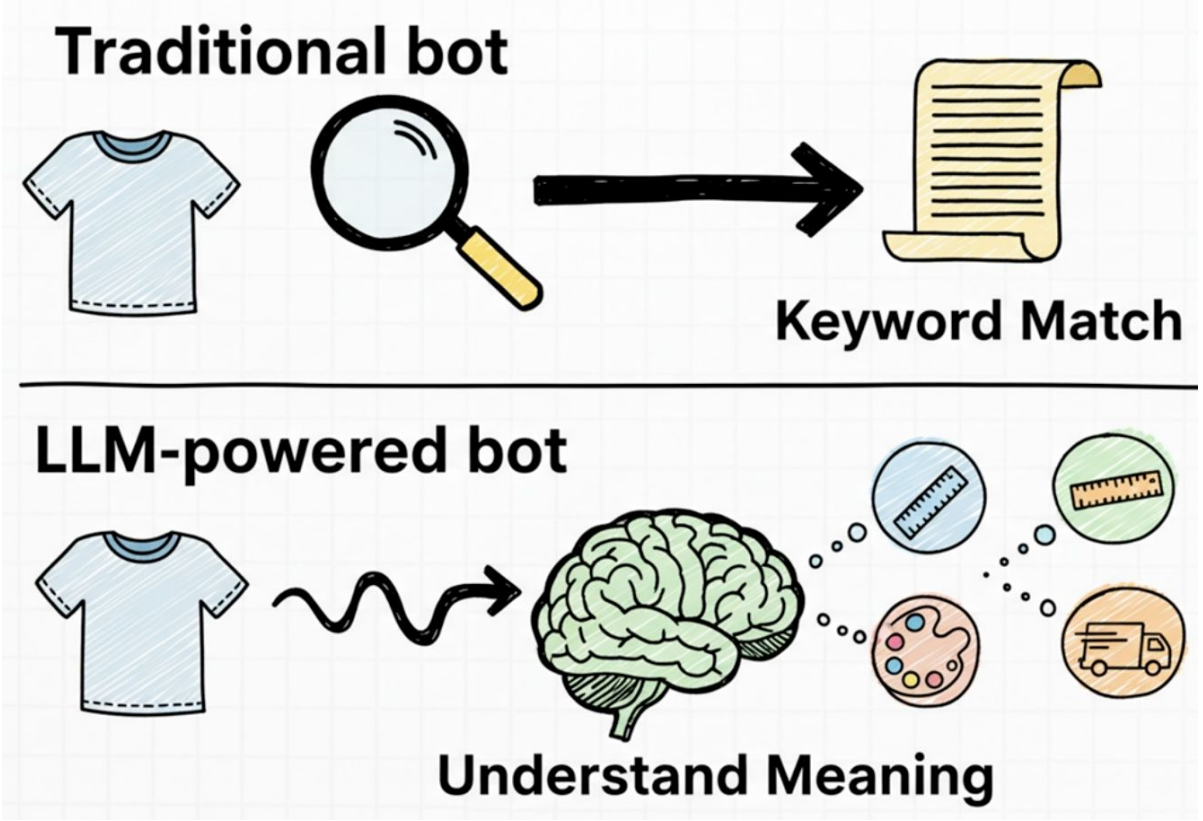
The Evolution

Traditional Chatbot

A scripted matching game. It scans for keywords ("shirt", "size") and fires a pre-written answer. It's fragile—a simple typo completely breaks it. It's like a basic search bar.

LLM-Powered Bot

They don't just see words; they understand the context and meaning behind the sentence. They handle slang, adapt naturally, and act like a real product expert who truly understands what you need.



Business Impact & Scale

80%
Seller Deflection Rate

+2.0%
Conversion Uplift

Performance Under Pressure

What does an 80% deflection rate actually mean? It means that **eight out of every ten customer problems get completely solved by the AI**. No human agent needed.

For Shopee, operating across 8 massive markets, this is AI in the wild. It sustains a **70% Buyer Satisfaction (CSAT)** score while handling millions of interactions daily.

Today's Agenda

1. The Landscape: E-Commerce Dilemma
2. **Core Architecture: 4 AI Pillars**
3. Training the Brain: SFT & Alignment
4. Conclusion & Key Takeaways

The Team of Specialists

1. Intent Router

Smart Switchboard

Multi-tier LLM dialog orchestration with lightweight models for routine intents and larger models for complex queries.

2. Product Expert

Deep Product Intelligence

The multimodal Q&A system for detailed, grounded answers.

3. Shopping Guide

Proactive Consultant

Function-call driven LLM recommendation engine with dynamic API invocation and real-time ranking.

4. Memory System

Curing Goldfish Memory

Cross-session memory capturing intents, entities, and behavioral patterns for long-term personalization.

A powerful assistant isn't one giant, all-knowing AI. It's an orchestra—a collection of different AI modules where each is designed to do one thing exceptionally well.

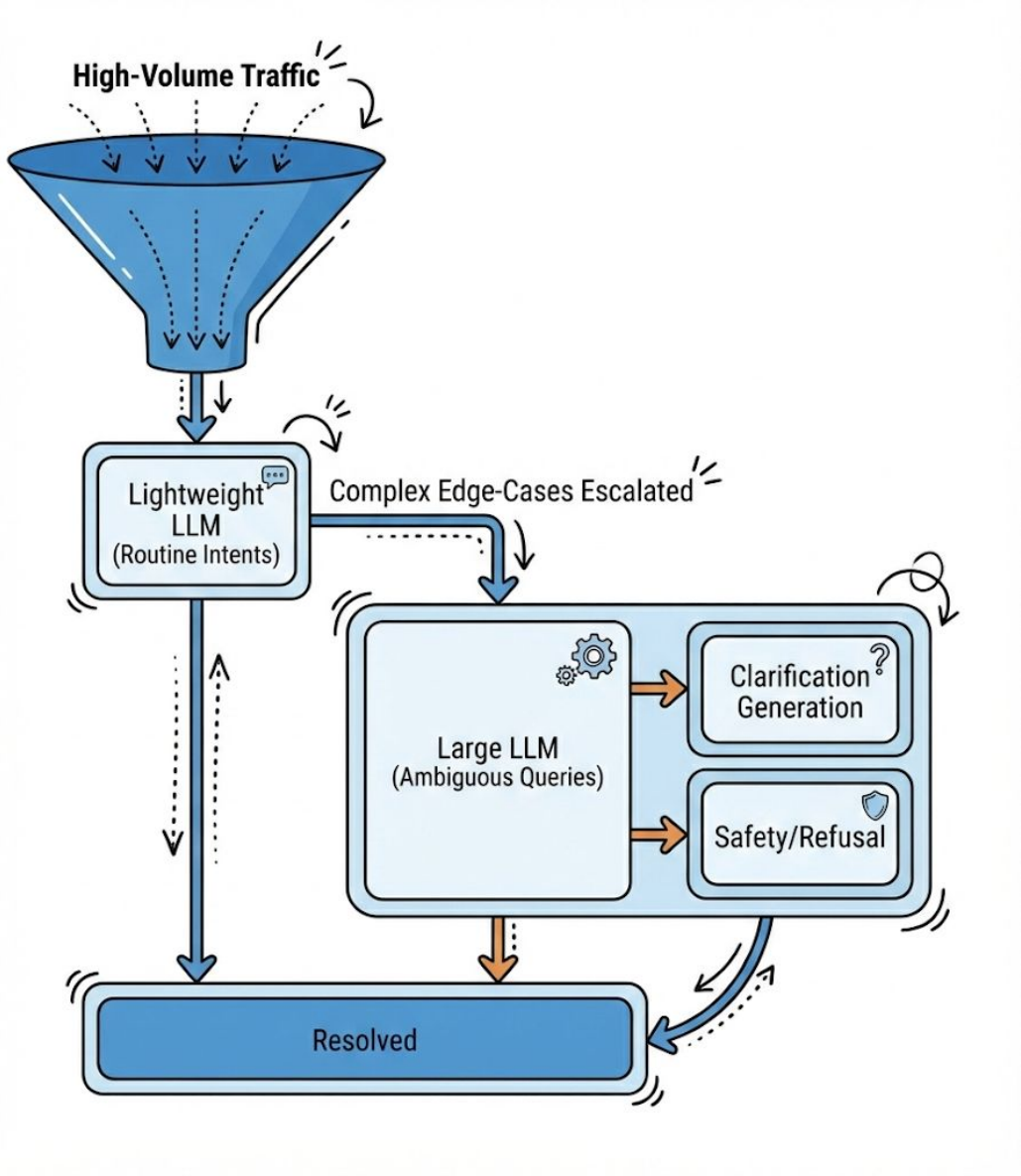
We've structured our architecture into 4 distinct pillars, moving from understanding the user, to answering them, guiding them, and remembering them.

1: Intent Router

The Scenario: Chaos to Order

Business Need: Millions of buyers input vague, multi-lingual, or slang-filled text. We must instantly know if they want a refund, a recommendation, or just shipping info.

Functional Module: The Switchboard.
The system's first job is simply to figure out what the person wants. It analyzes the text and chat history to route the request to the correct specialist module or safely escalate to a human agent.



1: Intent Router

Multi-Tier LLM Routing

We engineered a tiered framework. Lightweight models (fast, cheap) handle routine, deterministic intents.
Larger Language Models (LLMs) are dynamically triggered only to resolve ambiguous, multi-intent, or emotionally complex queries, parsing through regional slang flawlessly.

Clarification & Safety

Being smart means knowing what *not* to answer.

We built robust modules that proactively generate **Clarification Requests** when users are too vague, and **Safety Refusals** to elegantly decline off-topic or malicious inputs, protecting brand integrity. This directly ensures our **80% seller deflection rate**.

2: Product Expert

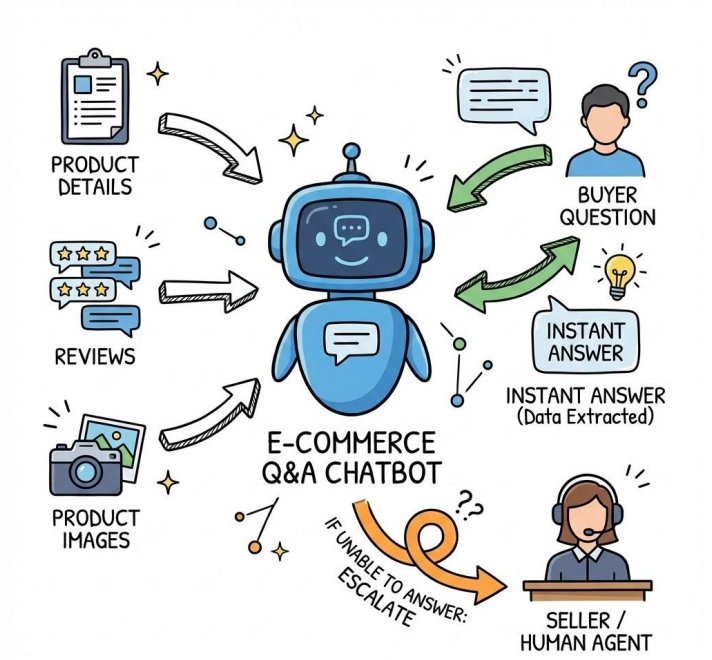
The Scenario: Deep Product Intelligence

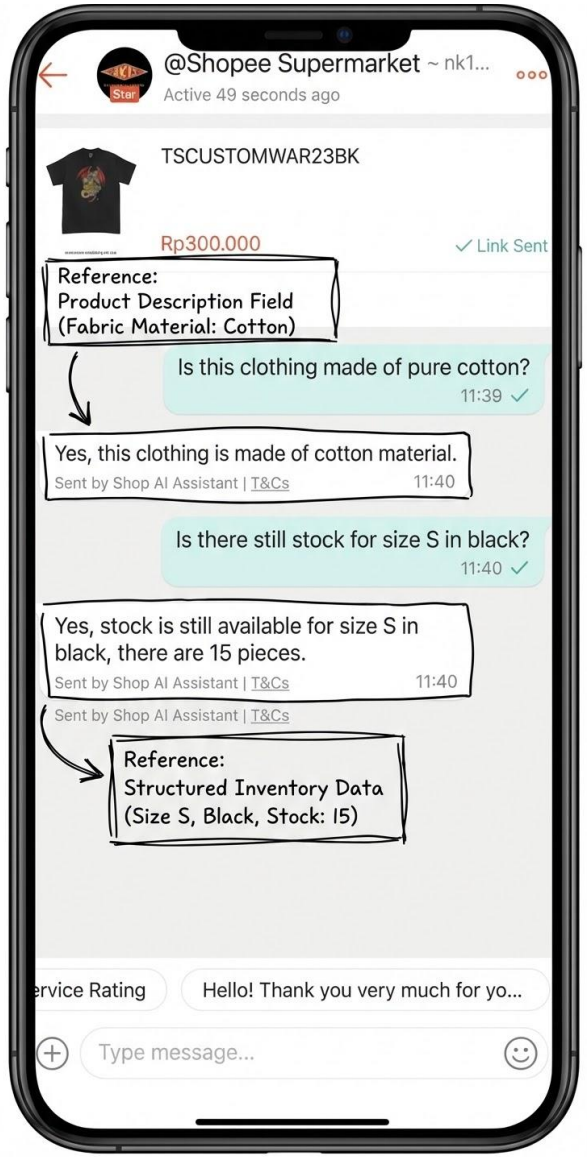
Business Need: A buyer asks, *"Is this shirt free size?"* or *"Does thi fit a 15-inch laptop?"* Generic answers kill the sale.

Deep Data Fusion: Synergizes unstructured manuals and technical parameters with real-world user data and FAQ libraries.

Fact Anchoring: Guarantees source-backed accuracy with zerc tolerance for "hallucinations" or false promises.

Intelligent Boundary Management: Automatically detects out-of-scope queries and triggers a smooth transition to live support.





2: Product Expert

Multimodal Product Q&A

Text alone isn't enough. We implemented **Multimodal Product Q&A**.

The LLM simultaneously analyzes unstructured text (descriptions), structured attributes (JSON data), user reviews, and product images to formulate a comprehensive answer that a human seller would give.

Hybrid Search & Fusion

To feed the LLM, we use a high-throughput **RAG Pipeline**.

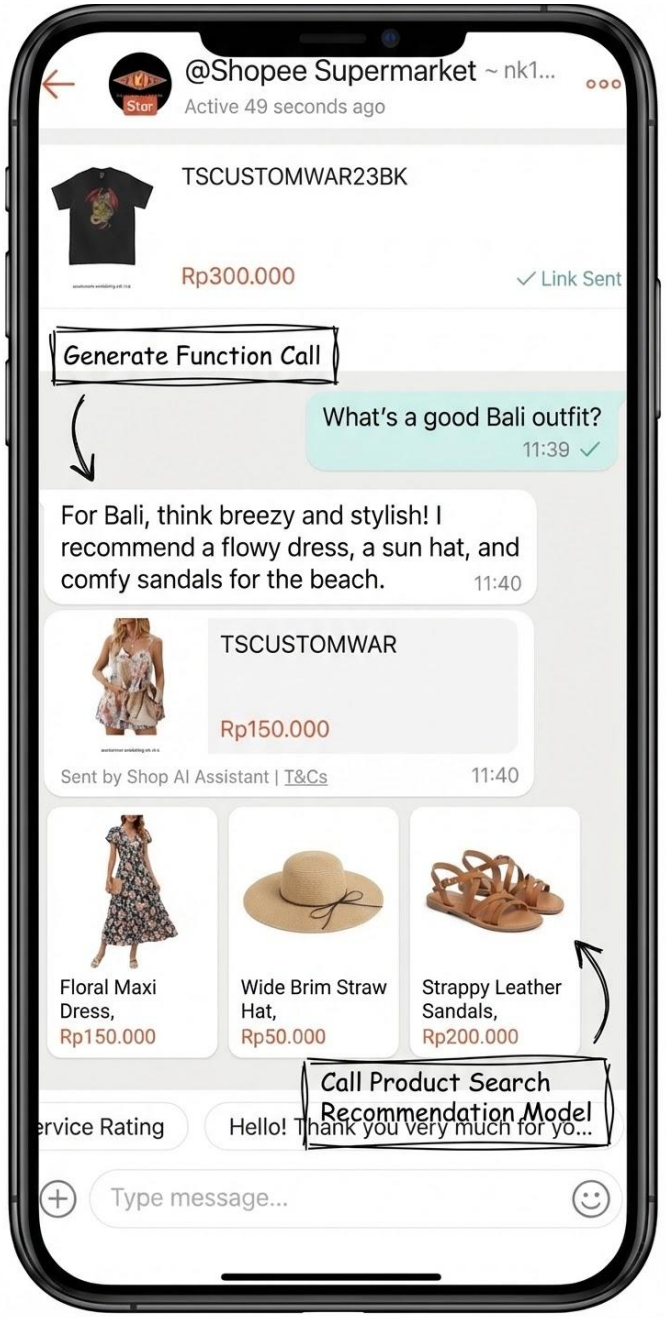
We combine dense vector search (FAISS) with sparse keyword matching to retrieve exact specs. We then use Context Fusion techniques to blend these multiple modalities into a single, hallucination-free prompt.

3: Shopping Guide

The Scenario: Vague Desires

Business Need: Users rarely know exactly what they want. They type *"I need an outfit for Bali"*. Traditional search fails completely here.

Functional Module: Active Discovery.
The Personal Shopper module transforms the bot from a reactive FAQ responder into a proactive consultant. It guides the user, narrows down choices, and actively pushes personalized item cards to drive conversion.



3: Shopping Guide

Function Call & Requirement Extraction

How does "Outfit for Bali" become a product? We use **LLM Function Calling**.

The LLM parses the natural language, extracts implicit requirements (e.g., Category: Summer Wear, Fabric: Breathable), and automatically structures them into JSON parameters to query backend APIs.

Personalized Recommendation

We bridge the LLM with our legacy RecSys.

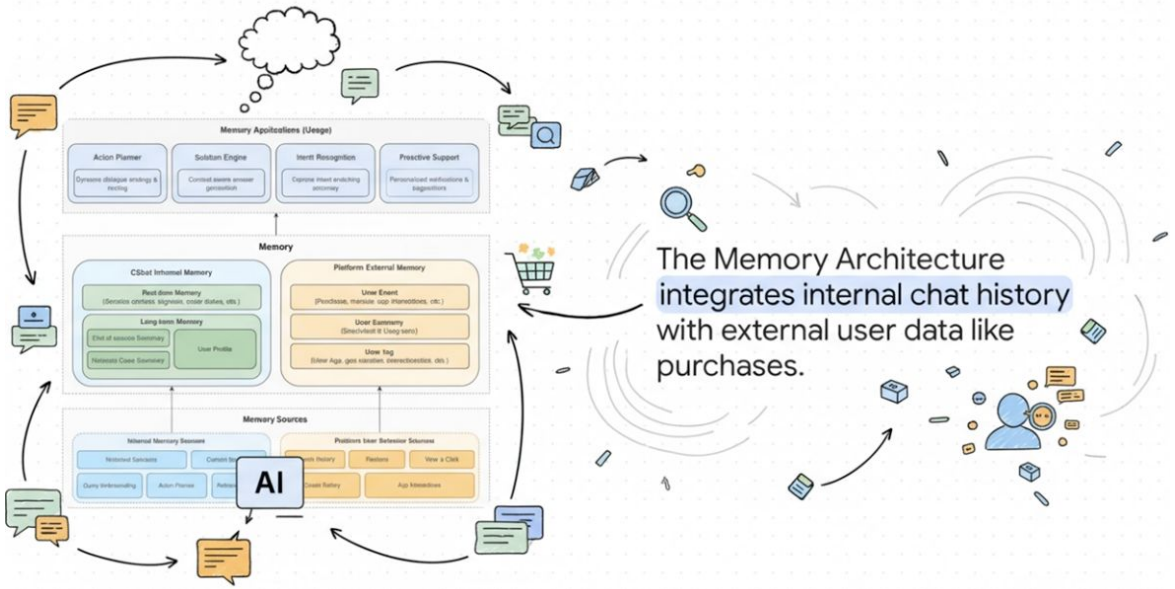
By executing these Function Calls against the Recommendation API, and using offline embedding refinement to match user profiles, the LLM serves items highly likely to convert, driving our **+2.0% Conversion Uplift**.

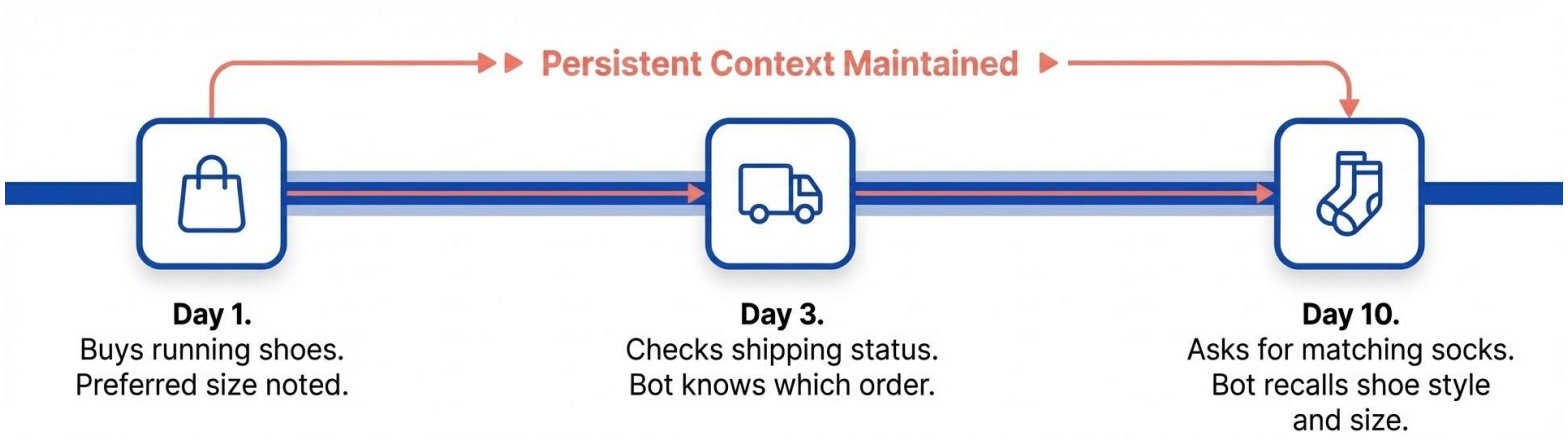
4: Memory System

The Scenario: The Frustrated User

Business Need: Users hate repeating themselves. A user returning with a recurring issue gets angry if treated like a stranger.

Functional Module: Curing Goldfish Memory.
The Memory Module tracks conversational state and historical behavior. It ensures the assistant remembers context within the current chat, and across days or weeks of platform usage.





4: Memory System

Dual-Layer Memory Architecture

Event Tracking: Individual sessions are synthesized into Topic-Based Threads, ensuring continuity by tracking ongoing projects and recurring themes across multiple conversations.

Dynamic User Profiling: Long-term data is distilled into an Evolving Persona Profile, continuously updating your core preferences, expertise, and communication style into a persistent identity ledger.

Cross-Session State Injection

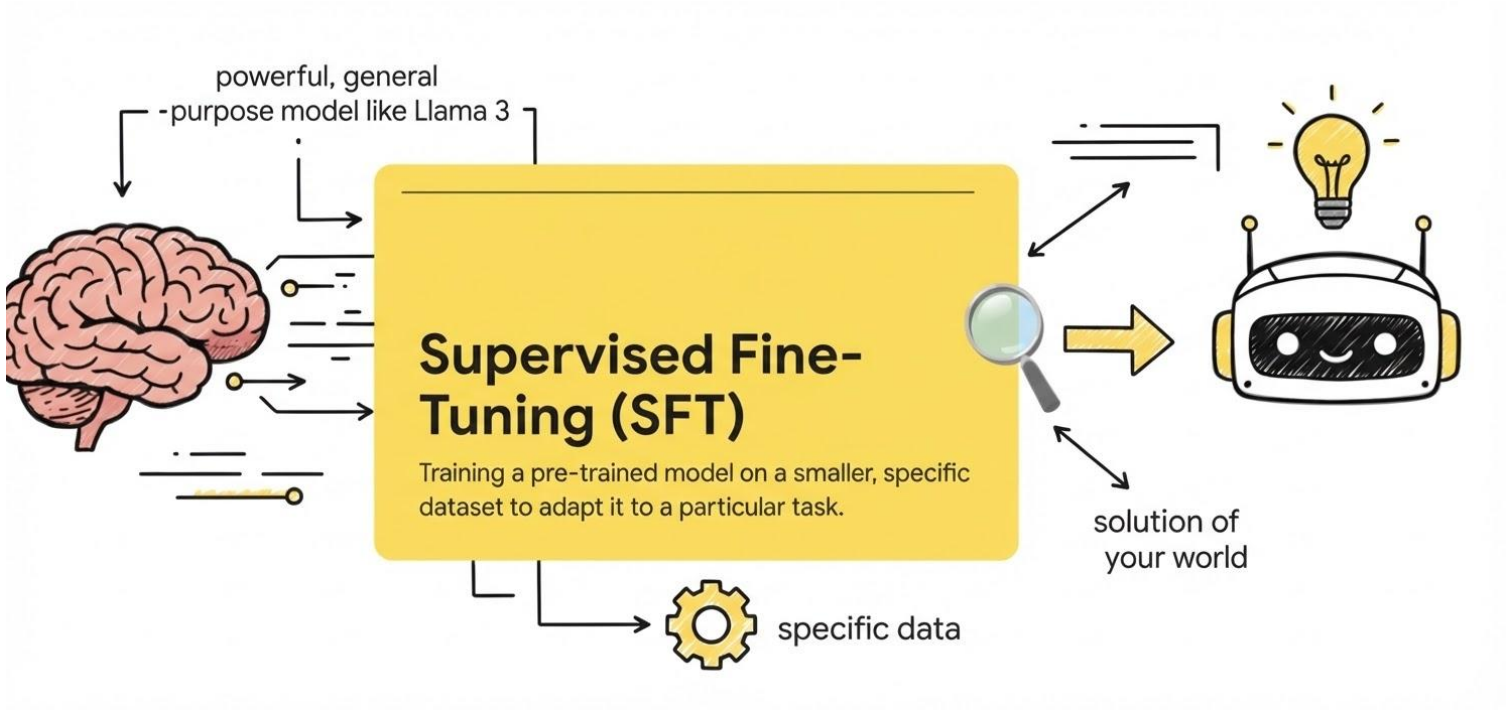
We don't just store data; we actively fuse it. When a user returns, their personalization is injected straight into the LLM prompt.

This allows the AI to immediately recognize user, skip redundant questions, and tailor responses natively, drastically improving satisfaction scores.

Today's Agenda

1. The Landscape: E-Commerce Dilemma
2. Core Architecture: 4 AI Pillars
3. **Training the Brain: SFT & Alignment**
4. Conclusion & Key Takeaways

The Brilliant New Employee (SFT)



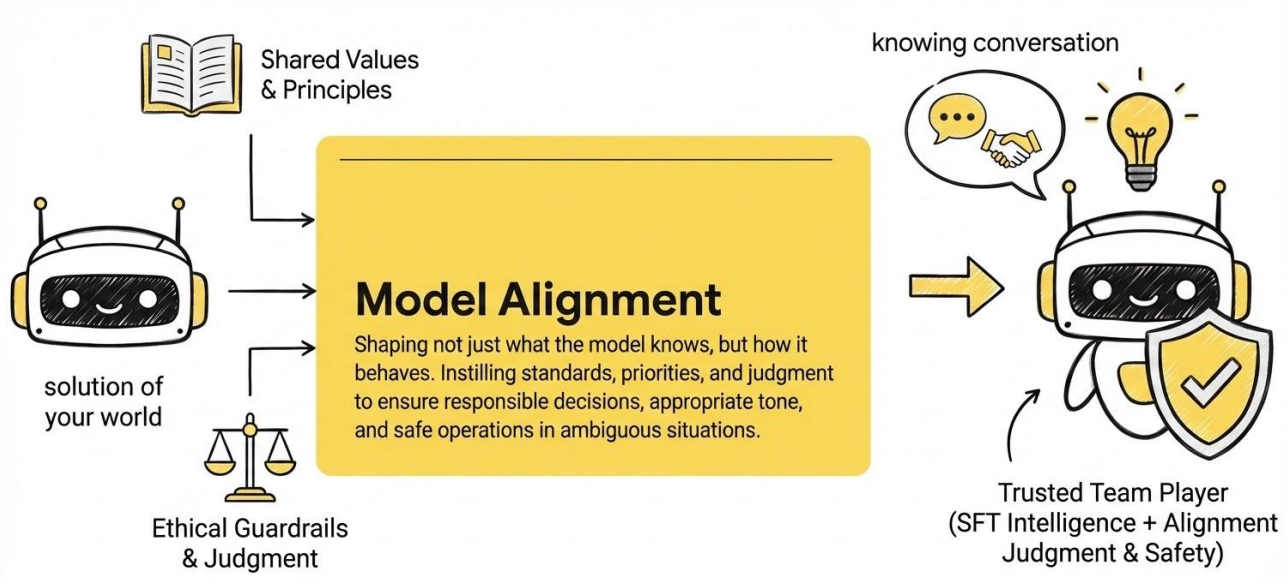
A base AI model has raw potential but knows nothing about your business. Through Supervised Fine-Tuning (SFT) on our specific high-quality domain data, we taught it the unique language, problems, and solutions of our world—yielding a massive leap in task accuracy.

Model Alignment

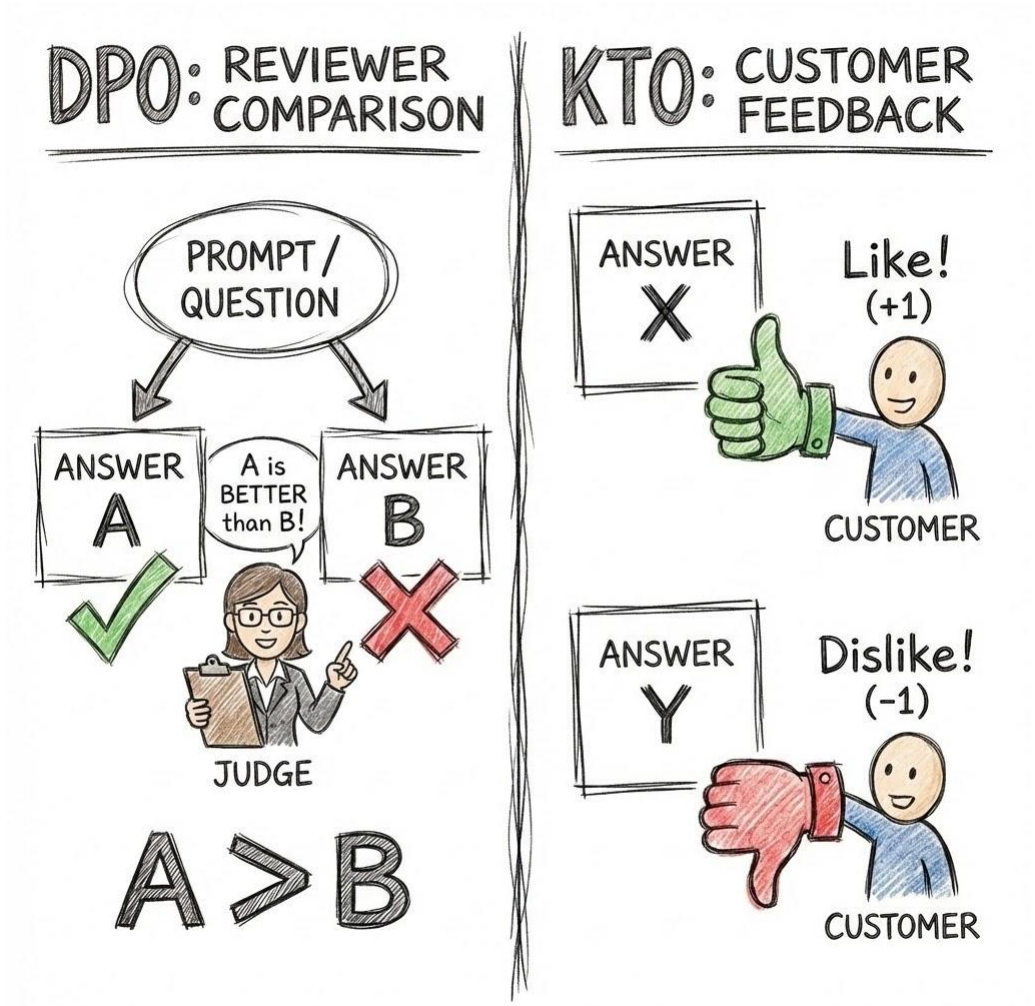
Safe & Helpful

Being accurate isn't enough. The model's answers also need to be genuinely helpful, incredibly safe, and aligned with what users actually want.

This is the next level of refinement. It ensures the AI doesn't just spit out technically correct but useless data, but instead acts with empathy and strategic helpfulness, matching the tone of top human sellers.



The Engineering Trade Off: DPO vs KTO



Optimizing Preferences

DPO (Direct Preference Optimization): Showing the AI two answers and declaring *"this is better than that."* Yields higher performance but requires complex, paired data.

KTO (Kahneman-Tversky Optimization): Simpler. Labeling answers purely as *good or bad*.

It's a classic engineering tradeoff: balancing the push for maximum generative performance against the effort of data annotation.

Today's Agenda

1. The Landscape: E-Commerce Dilemma
2. Core Architecture: 4 AI Pillars
3. Training the Brain: SFT & Alignment
4. **Conclusion & Key Takeaways**

"Our mission is to transform passive e-commerce browsing into an active, consultative conversational journey."

Key Takeaways

1. Proactive vs Reactive

LLMs fundamentally transform chatbots. They elevate the system from a fragile, keyword-matching FAQ bot into a proactive, consultative sales associate capable of driving discovery.

2. Orchestration Over Monoliths

A production-grade AI isn't one giant model. It's an orchestrated symphony of specialized modules (Intent, Q&A, Guide, Memory) that balance latency, cost, and extreme accuracy.

3. Deep Engineering Drives ROI

Advanced techniques—like Multimodal PQA, Function Calling for RecSys, and SFT—are what yield real business impact: 80% Seller Deflection Rate, 70% Buyer Satisfaction, and a +2.0% Conversion Uplift.

Thank You & Q&A

Our products and services differentiate us

Huiyu Chen | Senior Machine Learning Engineer @ Shopee